Accred Qual Assur (2014) 19:245–257
DOI 10.1007/s00769-014-1061-8

REVIEW

Proficiency tests for laboratories: a systematic review

Filipe Medeiros de Albano · Carla Schwengber ten Caten

Received: 5 February 2014 / Accepted: 5 June 2014 / Published online: 6 July 2014

Abstract Laboratories analytical results must be reliable. Proficiency tests (PT) main objective is to provide independent demonstrations of laboratory competence. The demand for these activities is increasing on the world scenario, as well as its importance. The main objective of this study is to identify the important publications of the PT theme from 2005 to 2012 based on a systematic review procedure. The method proposed reached a total of 113 papers published in indexed journals and another 34 additional references, including standards, guidelines and recommendations of international or regional accreditation body cooperation, international standards organization and international metrological institutes. All selected references were clustered based on its approach, and then the main practices were presented. The approaches identified were related to performance assessment, calculation method for performance assessment, use of PT for validation and/or estimation of measurement uncertainty and management of PT. Results indicate some opportunities to develop researches, such as: project management related to PT, the importance of analyses of data probability distribution function when consensus value is used, criteria to select the parameter to homogeneity and stability tests and to explore the link between PT, method validation and measurement uncertainty, among others.

F. M. de Albano (✉) · C. S. ten Caten
Universidade Federal do Rio Grande do Sul (UFRGS),
Ave. Osvaldo Aranha DEPROT/PPGEP, 99, 5° Floor,
Porto Alegre, Rio Grande do Sul 90035-190, Brazil
e-mail: filipealbano@yahoo.com.br

C. S. ten Caten
e-mail: tencaten@producao.ufrgs.br

Keywords Proficiency test · Systematic review · Interlaboratory comparisons · Laboratories

Introduction

The requirements of the industries, the government and the service sector in relation to the quality of testing and calibration have expanded over the past few years. From this fact, the laboratories of different areas have come to fulfill with accreditation criteria.

Within the context of the qualification of laboratories, there are the requirements of ISO/IEC 17025 and ISO 15189, in which it is explicit that the laboratory should monitor the validity of tests and calibrations performed through a procedure of quality control. Such monitoring can be accomplished through participation in a proficiency test (PT) [1].

The PT are important programs that support the reliability of tests and calibrations. They are programs that compare results among a group of laboratories, with the goal of evaluating the technical competence for performing a method of testing or calibration [2]. After participating in a PT, the laboratory has evidence regarding its measurements, checking its proficiency.

In Brazil, the participation in PT is a prerequisite for requesting accreditation by National Institute of Metrology, Quality, and Technology (INMETRO). PT are needed in the routine entities seeking their qualification and recognition of third parties [3]. Laboratories have difficulty validating methods and evaluating measurement uncertainty and these activities can be supported by PT [4].

The PT are conducted through a system that aims to support the testing and calibration laboratories, ensuring the services offered and providing important information to

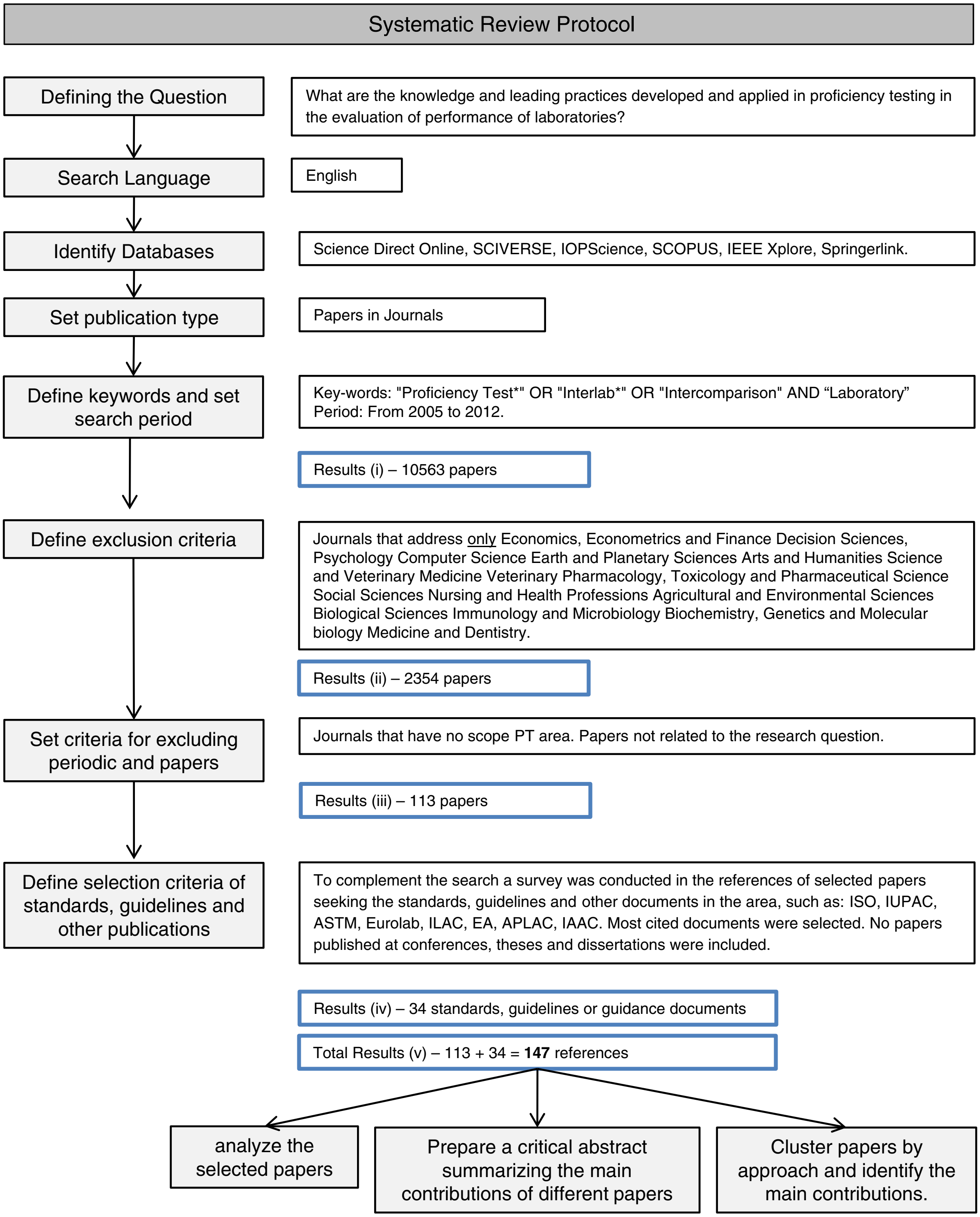
Systematic Review Protocol
Defining the Question
What are the knowledge and leading practices developed and applied in proficiency testing in the evaluation of performance of laboratories?
Search Language
English
Identify Databases
Science Direct Online, SCIVERSE, IOPScience, SCOPUS, IEEE Xplore, Springerlink.
Set publication type
Papers in Journals
Define keywords and set search period
Key-words: "Proficiency Test*" OR "Interlab*" OR "Intercomparison" AND "Laboratory" Period: From 2005 to 2012.
Results (i) – 10563 papers
Define exclusion criteria
Journals that address only Economics, Econometrics and Finance Decision Sciences, Psychology Computer Science Earth and Planetary Sciences Arts and Humanities Science and Veterinary Medicine Veterinary Pharmacology, Toxicology and Pharmaceutical Science Social Sciences Nursing and Health Professions Agricultural and Environmental Sciences Biological Sciences Immunology and Microbiology Biochemistry, Genetics and Molecular biology Medicine and Dentistry.
Results (ii) – 2354 papers
Set criteria for excluding periodic and papers
Journals that have no scope PT area. Papers not related to the research question.
Results (iii) – 113 papers
Define selection criteria of standards, guidelines and other publications
To complement the search a survey was conducted in the references of selected papers seeking the standards, guidelines and other documents in the area, such as: ISO, IUPAC, ASTM, Eurolab, ILAC, EA, APLAC, IAAC. Most cited documents were selected. No papers published at conferences, theses and dissertations were included.
Results (iv) – 34 standards, guidelines or guidance documents
Total Results (v) – 113 + 34 = 147 references
analyze the selected papers
Prepare a critical abstract summarizing the main contributions of different papers
Cluster papers by approach and identify the main contributions.

Fig. 1 Protocol of the systematic review

the company quality management [5]. Through PT, it is possible to evaluate the performance of laboratories for specific tests or measurements, identifying analytical problems, establish comparability of methods for testing or calibration, provide additional assurance to laboratory customers, enabling participants based on results of interlaboratory comparisons, validate the declared uncertainty and assign values to reference materials [6].

The comparison programs may vary according to the needs of the industry in which they are used, sample characteristics, methods in use and the number of participants. The nature of the test or measurement taken in PT defines the method of comparison of performance, which can be quantitative, qualitative or interpretive [6].

Within this context, we highlight the following research question: What are the main practices and knowledge developed and implemented in the performance assessment of laboratories in PT?

The purpose of this study is to identify and analyze the knowledge and leading practices developed and implemented in PT. Our specific goal is to identify the key trends in this area and the theoretical gaps in the development of PT. This article is structured in four sections: introduction, description of the protocol of the systematic review, analysis of results and conclusions.

Protocol of the systematic review

This is a research of applied nature and has goals of exploratory character and is dependent on knowledge of the primary sources consulted. The approach of this study is considered qualitative. The proposed logic to perform the systematic review is described in Fig. 1, which shows the method used to perform the search, critical appraisal and synthesis of the information selected.

The proposed method is based on the concepts presented by Akonbeng [7]. The systematic review was chosen for this study based on the statements made by the author cited above, who highlights the fact that this kind of work enables incorporating a larger number of contributions of relevant results, rather than just limiting the completion of some authors, allowing generalizability of the results.

The first stage of the protocol consisted on the elaboration of the research question which underlies the research proposal. The next step was to identify in what language the search would be performed, which was defined as only English. The survey was conducted in six scientific databases, where the initial focus of the search was papers published in indexed journals.

The keywords of the search were defined using Boolean logic, with applications such as OR, AND and *. The survey period was also limited between 2005 and 2012 (until June). The first result was 10563 papers. Subsequently areas that dealt only with specific matters, without addressing the research question and the theme of PT, were excluded. At this point, the number decreased to 2354 papers. As inclusion criteria for journals to prioritize the search, it was stipulated that those with the area of PT in its scope and papers that were directly related to the research topic would be considered, amounting to 125 studies selected. The analysis and selection of articles were conducted through a critical reading of their abstracts, and 12 papers were discarded because they were not directly related to the research question. Thus, 113 papers were selected.

The last step was a secondary search in the references of selected papers, identifying key standards, guides and recommendations from entities related to Metrology, Accreditation and Quality areas, in which 34 more references were added. Other details about the protocol of the systematic review are described in Fig. 1.

As Fig. 1 demonstrates, the implementation of the protocol generated a total of 113 articles and 34 additional references. The next section, which discusses the results of the study, presents a critical analysis of selected documents as well as the major theoretical and practical contributions identified, in order to develop considerations on the subject exposed.

Analysis of the selected documents

The analysis of the references researched was divided into two distinct parts: papers published in journals and standards and other documents in the field of PT. Following are the key concepts and practices identified.

Papers published in journals

We selected 113 papers published in journals that presented the topic PT and were directly related to the research. The main journals selected were: Accreditation and Quality Assurance (80), Flow Measurement and Instrumentation (2), IEEE Transactions on Instrumentation and Measurement (9), Measurement Transaction (11), and Metrologia/BIPM (11). The selected studies were critically analyzed and classified into four sub-areas, according to their approach and the application of the research conducted, namely: performance assessment in PT; calculation method for performance assessment in PT; use of PT for validation and/or estimation of measurement uncertainty, and management and improvements obtained in PT. There are some papers that can be related to more than one approach, but we analyzed them and classified in the approach that has more correlation to the article. This classification is presented in Table 1.

Papers about performance assessment

Several publications analyzed are connected to the use of PT for performance assessment of laboratories, where they are used to confirm modifications or improvements made in measuring methods, and may also be used to assess different measurement systems [9, 11, 13, 14, 23, 40, 49]. This PT is usually made with reference laboratories involved, which can come from National Metrology Institutes (NMI) [10]. Comparisons are also frequently made between NMI, called key comparison, which are relevant to ensure the measurements made by NMI are equivalent [11]. The purpose of key comparison is to support equivalence of measurements of NMI.

Table 1 Approaches and selected papers

Approach	Paper
Performance assessment in PT	[9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 21, 23, 29, 40, 49, 60, 63, 95, 96, 97, 104, 105, 106, 108, 109, 110, 111, 112, 113, 114, 115]
Calculation method for performance assessment in PT	[19, 22, 24, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 41, 42, 44, 45, 46, 50, 52, 54, 55, 56, 57, 58, 59, 61, 65, 66, 68, 69, 71, 73, 76, 78, 79, 81, 83, 85, 87, 89, 90, 91, 93, 98, 100, 103, 107, 116, 118]
Use of PT for validation and/or estimation of measurement uncertainty	[8, 20, 25, 38, 43, 53, 62, 64, 67, 70, 75, 77, 80]
Management of PT and Improvements obtained in PT	[47, 48, 51, 72, 74, 82, 84, 86, 88, 92, 94, 99, 101, 102, 117, 119, 120]

Comparisons with long rounds can use NMI reference laboratories and also pivot laboratories, which make intermediate measurements and are considered to be sub-references, and can participate in the stability study of the artifact [13].

Another common practice is to perform bilateral comparisons, generally made between two laboratories, where the one that has the best measurement capability, that is the lowest uncertainty, is designated as the Reference [15–18, 21]. Bilateral comparisons can be made with or without the presence of an NMI.

PT performance assessment also allows predictions concerning the analytical performance of laboratories in one country or a large organization. Research indicates that, in the field of microbiology, it was possible to assess the performance of Belgian and Canadian laboratories in a project for technical improvement of laboratories [60]. A similar approach was presented in comparisons made in other countries such as Croatia, Finland, France, Germany, Hungary, Russia, Slovenia, Spain and Switzerland, where one can have an overview of the participants and can assess the quality of results issued broadly, identifying regional deficiencies [66].

The performance evaluation of laboratories can also be accomplished through the use of Certified Reference Materials (CRM), with property values already known. In this case, as the number of participating laboratories increases, the cost of PT increases, since these materials are expensive [63]. Another possibility is to use a consensus value or historical value of other PT. This approach is also discussed in the next section.

Bilateral comparisons are more frequent in the area of calibration or physical measurement systems. In the testing area more specifically in the chemical and biological areas, the most frequent type of PT is the simultaneous samples, where there are rounds of comparison with several laboratories (usually more than 20 involved).

Papers about calculation method for PT performance assessment

Most of the papers analyzed discuss the statistical methods used to evaluate the performance of laboratories in PT. Surveys indicate that there is a reasonable harmonization in the use of indexes such as *z*-score and Normalized Error (E_n) but the procedure used to set the assigned value and the standard deviation (s) or uncertainty of reference are not harmonized [24, 41, 43, 46, 61, 68, 71, 73, 90, 98].

A result is considered satisfactory when the absolute value of the *z*-score is less than or equal to 2, questionable when it is between 2 and 3, and unsatisfactory when it is equal to or larger than 3. Already $|E_n|$ should be smaller than 1, so the results were satisfactory. The estimated reference standard deviation and measurement uncertainty need to be reliable. When they are not correctly estimated, the performance evaluation can be considered inconsistent [46].

There are different approaches to obtaining the assigned values in PT. The safest way is to obtain the value of a known sample, such as a CRM, or a reliable reference laboratory such as an NMI. Accredited laboratories could also be considered to be a reference, but for this, they should, in addition to accreditation, provide a suitable measurement capability (a reduced uncertainty) [41]. In the latter case, a prior demonstration of proficiency would also be advisable.

One of the common approaches in terms of calculation methods for PT performance assessment is the use of consensus value, calculated by classical or robust statistics. The reliability of the determination of the consensus value is relevant, since the mean, median or mode calculated will be designated as the reference value for PT. The estimated standard deviation also plays a key role in the evaluation of performance, so it must be assessed by the PT provider with caution [59, 61].

A study of PT providers from different European countries and the USA (in the health area with hemoglobin and leukocytes analyses) indicated that the method of calculation used for performance assessment does not have a standard [66]. The exclusion of outliers was performed by providers who participated in the survey, but using different procedures (for instance, in Russia values above $2s$ were considered outliers, in Finland, Spain, France, Hungary and Slovenia values above $3s$, and in Germany values deviating from the value of the median by more than 40 %). In this same survey, the designation of the assigned value was performed in different ways: in Germany and Slovenia reference labs were involved; in Croatia, France, Russia, Spain and Finland the mean value was applied, while in Belgium and Switzerland the median value; and in Hungary the result of a specialist laboratory was used as the assigned value. This demonstrates the lack of standardization of the different providers. The criterion used for satisfactory results was also variable. Half of the countries surveyed used a percentage deviation from the designated target value as criterion, which ranged between 3 % and 25 %. These values are usually stipulated by the legislation of those countries. The other providers work with the criterion based on the deviation of PT, and the range of satisfactory ranged from *s* to 2*s* [66]. Besides the differences observed, the decision about the PT Scheme design in performance evaluation is not a cultural issue, not even a regional issue. The decision is based on the PT Scheme provider, except in the case that designs are set by regulation.

Several studies indicate that the probability distribution of the PT data, when working with consensus value, should also be considered [30, 32, 50, 116]. Ideally, it should follow a Gaussian distribution, that is, symmetrical. If the associated probability distribution is not normal, the assessment by consensus value may be impaired (in the case of bimodal or asymmetric distributions, for instance) [30].

Another important issue is when the number of laboratories is reduced in a PT (<30, for example), because one should be more careful in performance assessment, since the reliability regarding the estimated reference standard deviation tends to decrease significantly [57]. Another case that deserves special attention is when the amounts of the analyses of interest are very low, because in this case the use of the standard deviation of consensus may not be the best alternative. Studies indicate that the proposal of Horwitz or the determination of deviations based on historical data of rounds, considering the mass fraction of the element analyzed, prove to be the most appropriate alternatives for the designation of reference standard deviation [68, 73, 89].

Researchers have also conducted simulations to verify the suitability of the use of consensus values of the PT through the Monte Carlo method [41, 79]. In these studies, it is clear that the concentration of the analyses, the method bias, the tendency of the laboratory and its repeatability can affect the consensus value. Even so, the approach of using the consensus value was considered adequate (considering the different simulated scenarios). It is worth highlight that studies comparing the use of consensus value with the use of CRM as a reference value were also conducted.

It was observed that performance assessment by E_n is more frequent in the calibration area. This index is the absolute value of the ratio between the difference of a value measured by a laboratory and a reference laboratory and the root of the quadratic sum of the expanded uncertainties (of the laboratory being assessed and of the reference laboratory). Usually, this index must be less than 1 to be satisfactory, but it is also possible that the evaluation criteria are less than 2 when working with a standard uncertainty [28, 39]. However, it is not possible to assess the performance of E_n only mathematically. This index is valid only if the uncertainty of the reference value is less than or equal to the uncertainty of the laboratory being assessed. Studies show that even labs with $E_n < 1$ still may have inadequate results compared to the others [59]. Other publications also comment on the necessary caution when comparing results with high uncertainty, which benefits the laboratories with a high random error. Moreover, it is commented on the problem of using only the *z*-score between laboratories in PT, which only evaluates the trueness of the laboratory, but does not account for its repeatability. Therefore, it is always necessary to consider uncertainty or its components in a consistent performance evaluation [65].

Other related approaches with methods for PT performance assessment can also be highlighted, such as: applications of a new statistical method ordinal analysis of variance (ORDANOVA) for interlaboratory comparisons with measurement or semi-quantitative (ordinal) and qualitative (binary) test results [78]; development of methods for quantitative analysis of PT [118], taking as an example the case of the Organization for the Prohibition of Chemical Weapons working with PT to verify that laboratories are able to identify prohibited chemical substances and hazardous samples [76]; using an average weighted by the uncertainty of the laboratories to create a consensus value and weighted averages with different criteria [26, 37]; application of ANOVA and ISO 5725 for performance assessment of PT participants [42]; use of PT participants' results to perform assessment of homogeneity and stability of the data rounds of comparison [44], among others.

Papers about method validation and estimation of measurement uncertainty

Papers classified within the approach of this section are linked to the use of PT in the validation of a method and to support the estimation of its uncertainty. It is possible to estimate measurement uncertainty through PT [20] using alternative approaches so that the comparison data can be combined with data from internal quality control of a laboratory, thus obtaining a combination of different sources of variability focusing on a reasonable estimate of the uncertainty of a trial. Different authors also comment about the use of PT in the validation of methods that have been modified from their original proposal and, after a comparison with other laboratories, may consider that the changes were consistent and appropriate [25]. This two uses of PT Schemes are not pointed in ISO/IEC 17043.

Still on the validation methods, it can be stated that PT results could be used as an alternative to meet certain requirements such as analytical precision, trueness and uncertainty [43]. Furthermore, the samples of PT could be used in internal quality control. This additional use of PT can help laboratories to reduce the financial impact of its quality assurance procedure [43].

The adequacy of performance assessment performed in a PT is linked to uncertainty of the assigned value. Within this context, it is possible to work with a "target uncertainty." The importance of implementing the "target measurement uncertainty" was indicated in different areas (testing and calibration). For a proper comparison, it was recommended that uncertainty target was at least three times less than the uncertainty of the participating laboratory [53]. This way, the laboratory can identify whether or not its uncertainty is appropriate [64].

Since the publication of the Guide to the Expression of Uncertainty in Measurement (GUM), many projects have been carried out to develop an alternative practice when it is technically or economically difficult to obtain a suitable mathematical model of the measurement [62]. Many laboratories are also reluctant to apply the law of propagation of uncertainty with its apparent mathematical complexity. These alternative practices can use the experimental data available from laboratories, such as repeatability, reproducibility, control charts, PT, among others. The only point to be noted in this approach is the fact that the standard uncertainty used based on the PT may be higher, because this proposal takes into account all the variability introduced by the different analytical methods. A more promising method for estimating uncertainty would be to use a combination of PT data and internal validation data of the method or quality control [62].

There is a mathematical model that was tested to estimate uncertainty of a laboratory, relating it to the standard deviation of the measurement and with the concentration of an analysis. This model was evaluated through a meta-analysis considering different PT, where its wide application was evident. The proposed mathematical function may be represented by the square root of the quadratic sum of α and $C{\cdot}\beta$, where C is the analyzed concentration [67]. The parameter α is connected to the detection limit of the method and β, to the relative accuracy of the method.

With these two parameters, a curve can be developed, where on the x-axis there is the mass fraction of the element that is being analyzed and on the y-axis, the standard deviation related to the concentration. Thus, it is possible to obtain the constants α and β of the mathematical model mentioned before and to obtain the standard deviation for reproducibility of the measurement system for any concentration value. This can be done with different analytical parameters. Obviously, a good estimate of model data depends on different concentrations of PT and preferably with a large number of participants. The reproducibility standard deviation is the major component of the standard uncertainty, from which the expanded uncertainty is obtained by multiplying with the coverage factor k; in most cases $k = 2$ is chosen for a confidence level of approximately 0.95 [67]. Other research on the same topic claim that this approach is useful and if applied appropriately makes available equations related to the performance of different analytical methods, besides the fact that the measurement uncertainty can be estimated for different concentrations [80]. It is worth highlight that these equations can be used to obtain an indication of the average quality of analytical results in a specific field and can be used by regulatory bodies to formulate legislation requirements according to the quality of existing measurement in the area [80].

Finally, other researchers indicate that the two most important concepts in metrology are certainly traceability of standards used and its measurement uncertainty, and its concepts are related to PT Schemes [75]. In areas such as chemistry and biology, many problems remain to be resolved to support international agreements related to these concepts. Therefore, NMI laboratories in these areas have developed strategies so that conclusions in PT are feasible and increasingly frequent [75], due to its importance and connection with traceability and uncertainty.

Papers about management and improvement of PT

The PT is developed by providers, who must also have proven their qualifications through an assessment of an accreditation body. These assessments are relatively recent, beginning through pilot programs, mainly in Europe, in 2005 [48, 88]. In Brazil, this activity became an official accreditation only in 2011, after the implementation of a pilot project by INMETRO.

International research conducted with 160 different providers from 32 countries show a strong tendency for accreditation of PT [47]. According to these surveys, it was found that this type of evaluation is based on various combinations of normative documents, which may illustrate a lack of harmonization of accreditation bodies. Furthermore, it was shown that some customers have an appeal to their suppliers to seek accreditation. However, among the providers consulted, less than half expect an improvement in their quality through accreditation and more than half expect a significant increase in their costs [47].

Another interesting approach is the possibility of organizing interlaboratory collaborative studies with a purpose of assessing the performance of the analytical test method and not only from laboratories [72]. Within this context, researchers recommend care in the management and conduct of a trial for purposes of performance assessment methods, as well as their statistical analysis. Issues such as the choice of participating laboratories and the designation of the assignment values are important. Therefore, it is clear that it is possible to establish a standard method for analysis through rounds of interlaboratory collaborative studies, with greater assurance that the developed method provides reproducibility in different operating conditions [72].

Requirements applicable to PT are similar to those considered in the production of reference materials [78]. The samples of PT should have a degree of homogeneity and stability for the purpose of identifying differences between the laboratories. Based on this logic, the process used to prepare the samples held by the provider must be appropriate and shall ensure the quality of the items that will be sent to laboratories in the comparison rounds [78]. Tests for homogeneity and stability are essential in this context.

Normally, PT are performed in rounds that occur during 1 year. Studies in the field of occupational medicine indicate that 28 % of PT run with 4 rounds per year [72]. Similar results were observed in hematology and microbiology, with a median of 3 rounds per year. The median of biochemistry was 6 rounds per year, where 33 % of the PT have intervals of 1 month. The number of samples per round varied between 1 (31 %) and over 20 (0.5 %), where most providers offer between 1 and 3 samples per round (83 %) [72].

The implementation of the PT has a wide area. Initially, they were most in demand in the area of calibration, being performed mainly by reference laboratories. The medical area also started with PT compulsory participation, due its importance. According to accreditation bodies, today the demand for PT in different areas is greater than its supply and availability. The expansion of PT is increasingly perceived in the field of chemical, biological, geological, agricultural tests and even in the veterinary area [51]. Nowadays, most of the PT done in the world is in medical areas.

Different international regulatory agencies also consider the PT as an appropriate way to ensure the reliability of laboratory results and, on several occasions, make participation in these activities compulsory [86]. Yet, research indicates that laboratories participating in PT over time tend to improve their results, as well as the providers improve the management and reliability of their programs [99, 101].

Providers also had to adapt and start work focused on better management of its activities, seeking compliance with standards such as ISO/IEC 17043 [117, 119, 120]. This standard addresses technical and managerial issues that should be followed by PT providers; however, it is still not compulsory to use in many countries. Meeting this standard, in an isolated way, when not assessed by a third part like an accreditation body, does not guarantee proper operation of the PT developed, since an adequate managerial capacity installed in companies and an appropriate technical knowledge on the subject are necessary.

Other selected references

The second stage of the systematic review focused on the pursuit of standards and guideline of renowned entities in the PT area. We selected the most-cited references in the articles that were considered in the previous step. Another 34 references were identified, from International Organization for Standardization (41.2 %), American Society for Testing and Materials (14.7 %), Asia Pacific Laboratory Accreditation Cooperation (14.7 %), International Union of Pure and Applied Chemistry (5.9 %), European Cooperation for Accreditation (5.9 %), European Federation of National Associations of Measurement, Testing and Analytical Laboratories (5.9 %), NORDTEST (2.9 %), Bureau International des Poids et Mesures (2.9 %), International Laboratory Accreditation Cooperation (2.9 %) and InterAmerican Accreditation Cooperation (2.9 %).

The selected references were classified into three approaches. The division performed is shown in Table 2. After the classification, a summary of the approach of these documents according to their classification is shown.

Definitions, management, operation and use of PT programs

Standards that address definitions of PT are mostly published by ISO. Some norms are for guidance [126, 128, 134–137], addressing specific PT in technical areas such as tissues, microbiology, petroleum products, among others. There are, in this group, standards that are used to accredit

Table 2 Approaches and other selected publications

Approach	Standards, guides and guideline documents
Definitions, management, operation and use of PT programs	[1, 2, 6, 121, 124, 126, 128, 129, 130, 134, 135, 136, 137, 144, 145, 146, 147, 148, 149, 150, 151],
Statistical methods for PT performance assessment	[122, 123, 125, 127, 131, 132, 133, 138, 139]
Use of PT to estimate measurement uncertainty	[127, 140, 141, 142, 143]

laboratories [1, 129] and that address PT in the field of quality assurance of testing or calibration.

Other standards are also used in the accreditation of reference material producers and providers of PT [6, 124], the latter of which establishes the technical and management requirements that must be followed to conduct a PT appropriately. Reference material producers and PT providers are different types of organizations, and they should not be confounded. Among the surveyed standards, ISO/IEC 17043 is the most complete and is used globally by different providers in different areas [6].

In this category, there are also standards [130] and other documents published by organizations that establish major policies for the accreditation process for laboratories and providers PT [144–146]. These documents establish the minimum frequency of participation in PT, which should be the policies of the bodies to assess inadequate results obtained in PT and how these factors may influence an accreditation process.

Statistical methods for PT performance assessment

Several standards and guideline documents have different approaches to PT performance assessment [122, 123, 125, 127, 131–133, 138, 139]. Most documents converge in the use of the same indicators for performance rating, the most common being the z-score (and its variations as Z'-score, zeta-score, etc.) and E_n. However, the method of calculation or estimation of reference values shows much divergence and relative lack of standardization. The standards usually present examples of the application of its procedures to set the assigned values, but they are general. It is common to need a “fit for purpose” in each specific PT Scheme developed.

Most documents propose the evaluation of repeatability, reproducibility and accuracy of the results of the participating laboratories in comparisons, but in a general way. Still, regarding the tests of homogeneity and stability of the items that are compared (samples or artifacts), we emphasize that the references do not provide details regarding how many analyses/parameters should be selected to consider testing representative and consistent. The documents cited in this section do not address in detail the influence that the probability distribution of the data may have on the results of PT.

Use of PT to estimate measurement uncertainty

Documents for estimation of the measurement uncertainty were also frequently referenced in the articles selected in this systematic review. Neither is focused only on PT, since they address methods for estimation of uncertainty in testing or calibration [140–143].

Furthermore, some documents suggest alternative approaches to calculating uncertainty, considering the results of PT [141, 142]. These approaches should be selected carefully, as the result of uncertainty can be strongly influenced by the performance of the participants of the comparison. Still, these alternative approaches are recommended when there is little information on the sources of variation of the method or when getting values associated with measurement accuracy is complex.

Regarding to uncertainty measurement of the assigned values, we think that is a point to be improved. The standard ISO13528 gives a very simple approach to establishing the uncertainty of the assigned value when the provider uses consensus value. In this case, the uncertainty can vary drastically according to the number of laboratories that are in the round [127].

Identification of gaps to be exploited

We can see the importance of the topic and the increasing demand for participation in PT, whether it is required by the government, accreditation and conformity assessment bodies. Due to the numerous areas of laboratories, providers are not yet prepared to meet all existing demands. Still, there is a perceived need for the structuring of these organizations in terms of obtaining adequate standards in the area and agile management to meet the market demands. Several PT are developed in different countries and different areas, but approaches that assist management of providers with a view of projects were not found in the sources researched. The main area of the reference standard, ISO/IEC 17043, also does not address the issue of development and management of PT with the project vision. It is likely that this is an issue to be explored. This standard does not consider areas such as risk, costs, strategy and time management, i.e., which are typical from project management knowledge, and could be useful in PT Schemes.

Although publications related to the topic often address the link between method validation, measurement uncertainty and PT, it is clear that there is not a document that presents a logical interface between these themes. This ends up creating doubts and does not always clear up what the actual intended use of PT is.

Another important issue, discussed by different researchers, is the impact that the probability distribution of the data can have on performance assessment. These issues, in most cases, are not considered by the providers and may have a high impact on the statistical treatment of data, especially when working with consensus value (with references generated with data from the participants of the

PT). Still, the standards of the area neither report details on this fact nor report procedures for assessing probability distributions obtained in PT.

It is noteworthy that a factor cited in different studies was the homogeneity and stability of the samples prepared in PT and the need for ensuring this point to increase confidence in the round of comparison. However, the standards and publications do not make clear what the criteria are for selection of parameters for these tests should be, as well as how many parameters would be representative for an adequate test of homogeneity and stability of the samples. This fact deserves attention, since a false sense of homogeneity or stability may compromise the trust of a PT.

Finally, the ISO Standards related to PT are, sometimes, general and not specific, because there are an enormous variety of measurement fields, national regulations and "fit for purpose" needs—one laboratory's needs for accuracy and precision are not always the same as another's.

Conclusions

This study presented a systematic review that covered the period from 2005 to 2012 (June) considering publications related to the theme PT. A total of 147 references were selected, including articles, standards and guideline documents.

Thus, it is considered that the research objective was achieved, since we analyzed the expertise and main practices related to the theme PT in the research sources listed above. These shortcomings were raised as follows: management of PT projects; analysis of the link between validation, PT and measurement uncertainty; preliminary evaluation of the probability distribution of the data from PT; selection of variables for testing homogeneity and stability. The shortcomings are not limited to these topics, though this analysis is based on the perception of the main factors analyzed. In future, researches or reviews about this theme are advisable to include published PT reports offered by international cooperation's (for example IMEP and APLAC) and private schemes that are offered internationally.

References

1. ISO/IEC 17025 (2005) General requirements for the competence of testing and calibration laboratories
2. APLAC (2008) PT-02: testing—interlaboratory comparisons. Issue No. 6
3. Tholen D (2011) Metrology in service of society: the role of proficiency testing. Accred Qual Assur 16:603–605. doi:10.1007/s00769-011-0836-4
4. Rosa D, Pizzolato M, Albano FM, Sartori T (2011) Analysis of non conformities of testing and calibration laboratory. In: Metrologia 2011—Metrology as factor of quality, innovation and competitiveness. SBM, Rio de Janeiro, Brazil. ISBM 978-85-86920-08-0
5. Howerton D, Krolak J, Manasterski A, Handsfield J (2010) Proficiency testing performance in US laboratories. Arch Pathol Lab Med 134:751–758
6. ISO/IEC 17043 (2010) Conformity assessment—general requirements for proficiency testing
7. Akobeng AK (2005) Understanding systematic reviews and meta-analysis. Arch Dis Child 90:845–848. doi:10.1136/adc.2004.058230
8. Choi HM, Park K, Youn KO, Choi YM (2010) Uncertainty evaluation procedure and intercomparison of bell provers as a calibration system for gas flow meters. Flow Meas Instrum 21:448–496. doi:10.1016/j.flowmeasinst.2010.07.002
9. Standiford DM, Lee M (2012) Inter-laboratory comparison results for Coriolis mass flowmeter calibration facilities. Flow Meas Instrum. doi:10.1016/j.flowmeasinst.2011.12.004
10. Hamilton CA, Kupferman SL, Salazar MT, Deaver D, Wood BM (2005) Interlaboratory comparison at 10 V DC. IEEE Trans Instrum Meas. doi:10.1109/TIM.2004.838134
11. So E, Angelo D, Tsuchiyama T, Tadokoro T, Waltrip BC, Nelson TL (2005) Intercomparison of calibration systems for ac shunts up to audio frequencies. IEEE Trans Instrum Meas. doi:10.1109/TIM.2004.843418
12. Bieler M, Seitz S, Spitzer M, Hein G, Pierz K, Siegner U et al (2007) Rise-time calibration of 50-ghz sampling oscilloscopes: intercomparison between PTB and NPL. IEEE Trans Instrum Meas. doi:10.1109/TIM.2007.890797
13. Tang Y, Hamilton CA, Deaver D, Parks H, Wood BM (2007) The seventh intercomparison of Josephson voltage standards in North America. IEEE Trans Instrum Meas. doi:10.1109/TIM.2007.890797
14. Jung JK, So E, Park YT, Kim M (2009) KRISS-NRC intercomparisons of calibration systems for instrument transformers with many different ratios at power frequency. IEEE Trans Instrum Meas. doi:10.1109/TIM.2008.2006966
15. Callegaro L, D'Elia V, Bohǎcek J (2009) Four-terminal-pair inductance comparison between INRIM and CTU. IEEE Trans Instrum Meas. doi:10.1109/TIM.2008.928402
16. Tang Y, Wood BM, Hamilton CA (2009) A two-way josephson voltage standard comparison between NIST and NRC. IEEE Trans Instrum Meas. doi:10.1109/TIM.2008.2006960
17. Landim RP, Tang Y, Afonso E, Ferreira V (2011) Josephson voltage standard comparison between NIST and INMETRO as a link to BIPM. IEEE Trans Instrum Meas. doi:10.1109/TIM.2010.2099370
18. Kang J, Kang N, Gentle DG, MacReynolds K, Francis MH (2011) Intercomparison of standard gain horn antennas at w-band. IEEE Trans Instrum Meas. doi:10.1109/TIM.2010.2103413
19. Pavese F (2006) A metrologist viewpoint on some statistical issues concerning the comparison of non-repeated measurement data, namely MRA key comparisons. Measurement 39:821–828. doi:10.1016/j.measurement.2006.05.007
20. Désenfant M, Priel M (2006) Road map for measurement uncertainty evaluation. Measurement 39:841–848. doi:10.1016/j.measurement.2006.04.008
21. Chung JW, Borys M, Firlus M, Lee WG, Schwartz R (2007) Bilateral comparison of buoyancy artefacts between PTB and KRISS. Measurement 40:774–778. doi:10.1016/j.measurement.2006.08.003
22. Ferrero C, Casaril M (2009) Proficiency testing programs to improve traceability in chemical analysis. Measurement 42:1502–1509. doi:10.1016/j.measurement.2009.07.005
23. Galliana F, Capra PP, Gasparotto E (2009) Report of the Italian inter-laboratories comparison of high dc resistance on the calibration of a 10 MX and a 1 GX resistors. Measurement 42:1532–1540. doi:10.1016/j.measurement.2009.07.014

24. Ferrero C (2009) The equivalence in accreditation: robust or standard statistical analysis of interlaboratory comparisons. Measurement 42:1541–1545. doi:10.1016/j.measurement.2009.07.011
25. Astrua M, Pisani M (2009) Validation of a novel technique for the measurement of the refractive index of a prism by means of interlaboratory comparison. Measurement 42:1546–1549. doi:10.1016/j.measurement.2009.08.012
26. Brunetti L, Oberto L, Sellone M, Terzi P (2009) Establishing reference value in high frequency power comparisons. Measurement 42:1318–1323. doi:10.1016/j.measurement.2008.08.008
27. Bergoglio M, Malengo A, Mari D (2011) Analysis of interlaboratory comparisons affected by correlations of the reference standards and drift of the travelling standards. Measurement 44:1461–1467. doi:10.1016/j.measurement.2011.05.012
28. Cascetta F, Comazzi M, Musto M, Rotondo G (2012) An experimental intercomparison of gas meter calibrations. Meas Trans. doi:10.1016/j.measurement.2012.04.020
29. Galliana F, Capra PP, Gasparotto E (2012) Inter-laboratories comparison at 100 GX and 1 TX level to evaluate the traceability transfer from INRIM in the field of high dc resistance. Measurement 45:615–621. doi:10.1016/j.measurement.2011.10.028
30. Willink R (2006) Forming a comparison reference value, from different distributions of belief. Metrologia. doi:10.1088/0026-1394/43/1/002
31. Hornikova A, Zhang NF (2006) The relation between the En values including covariance and the ‘exclusive’ En statistic. Metrologia. doi:10.1088/0026-1394/43/1/N01
32. Lira I (2007) Combining inconsistent data from interlaboratory comparisons. Metrologia. doi:10.1088/0026-1394/44/5/019
33. Kacker RN, Forbes A, Kessel R, Sommer K (2008) Bayesian posterior predictive p-value of statistical consistency in interlaboratory evaluations. Metrologia 44. doi:10.1088/0026-1394/45/5/004
34. Kacker RN, Forbes A, Kessel R, Sommer K (2008) Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations. Metrologia 45:257–264. doi:10.1088/0026-1394/45/3/001
35. Khlevnoy B, Sapritsky V, Rougie B, Gibson C, Yoon H, Gaertner A et al (2009) CCPR-S1 Supplementary comparison for spectral radiance in the range of 220 nm to 2500 nm. Metrologia 46:174–180. doi:10.1088/0026-1394/46/4/S08
36. Zhang W, Zhang NF, Liu H (2009) A generalized method for the multiple artefacts problem in interlaboratory comparisons with linear trends. Metrologia 46:345–350. doi:10.1088/0026-1394/46/3/024
37. Rukhin AL (2009) Weighted means statistics in interlaboratory studies. Metrologia 46:323–331. doi:10.1088/0026-1394/46/3/021
38. Arp U, Klein R, Li Z, Paustian W, Richter M, Shaw PS et al (2011) Synchrotron radiation-based bilateral intercomparison of ultraviolet source calibrations. Metrologia 48:261–267. doi:10.1088/0026-1394/48/5/004
39. Fellmuth B, Wolber L, Head DI, Hermier Y, Hill KD, Nakano T, Pavese F et al (2012) Investigation of low-temperature fixed points by an international star intercomparison of sealed triple-point cells. Metrologia 49:257–265. doi:10.1088/0026-1394/49/3/257
40. Kim M, Pratt JR, Brand U, Jones CW (2012) Report on the first international comparison of small force facilities: a pilot study at the micronewton level. Metrologia 49:70–81. doi:10.1088/0026-1394/49/1/011
41. Wong SK (2006) A comparison of performance statistics for proficiency testing programmes. Accred Qual Assur 12:59–66. doi:10.1007/s00769-006-0233-6
42. Svegl F, Strupi JS, Svegl IG (2006) Proficiency testing of chloride content in different types of Portland cement. Accred Qual Assur 11:414–421. doi:10.1007/s00769-006-0162-4
43. Detaille R, Maetz P (2006) Practical uses of proficiency testing as valuable tools for validation and performance assessment in environmental analysis. Accred Qual Assur 11:408–413. doi:10.1007/s00769-006-0156-2
44. Tholen D, Chappel R, Francis B (2006) Use of participant EQA results to assess sample homogeneity and stability for qualitative analytes. Accred Qual Assur 11:400–407. doi:10.1007/s00769-006-0167-z
45. Thompson M, Ellison SRL (2006) Fitness for purpose—the integrating theme of the revised harmonised protocol for proficiency testing in analytical chemistry laboratories. Accred Qual Assur 11:373–378. doi:10.1007/s00769-006-0137-5
46. Boley N, van der Veen AMH, Robouch P, Golze M, van de Kreeke J et al (2006) Comparability of PT schemes—what did we learn from COEPT? Accred Qual Assur 11:391–399. doi:10.1007/s00769-006-0106-z
47. Schmidt A, Örnemark U, Golze M, Henriksen GM (2006) Surveys on the accreditation of providers of proficiency testing and external quality assessment schemes. Accred Qual Assur 11:379–384. doi:10.1007/s00769-006-0110-3
48. Monnery K (2006) Experiences from the assessment of proficiency testing providers. Accred Qual Assur 11:385–387. doi:10.1007/s00769-006-0138-4
49. de Leer EWB (2006) The metrological infrastructure for gas analysis. Accred Qual Assur 11:175–181. doi:10.1007/s00769-006-0123-y
50. Kuselman I (2006) Nonparametric assessment of comparability of analytical results obtained in proficiency testing based on a metrological approach. Accred Qual Assur 10:659–663. doi:10.1007/s00769-006-0092-1
51. Colling A, Jeggo M, Louvandini H, Lelenta M, Robinson M (2007) The establishment of quality systems in veterinary diagnostic testing laboratories in developing countries: experiences with the FAO/IAEA External Quality Assurance Programme. Accred Qual Assur 13:33–45. doi:10.1007/s00769-007-0329-7
52. Rienitz O, Schiel D, Güttler B, Koch M, Borchers U (2007) A convenient and economic approach to achieve SI-traceable reference values to be used in drinking-water interlaboratory comparisons. Accred Qual Assur 12:615–622. doi:10.1007/s00769-007-0320-3
53. Serapinas P (2007) Approaching target uncertainty in proficiency testing schemes: experience in the field of water measurement. Accred Qual Assur 12:569–574. doi:10.1007/s00769-007-0310-5
54. Kuselman I, Belli M, Ellison SLR, Fajgelj A, Sansone U, Wegscheider W (2007) Comparability and compatibility of proficiency testing results in schemes with a limited number of participants. Accred Qual Assur 12:563–567. doi:10.1007/s00769-007-0309-y
55. Polkowska-Motrenko H, Rossbach M (2007) Preparation and characterization of a proficiency test material “Polish mushrooms” supported by the International Atomic Energy Agency (IAEA), Vienna. Accred Qual Assur 12:343–350. doi:10.1007/s00769-007-0293-2
56. Suchánek M, Robouch P, Aregbe Y, Herrero A, Taylor P (2007) Accred Qual Assur 12:399–403. doi:10.1007/s00769-007-0297-y
57. Belli M, Ellison SLR, Fajgelj A, Kuselman I, Sansone U, Wegscheider W (2007) Implementation of proficiency testing schemes for a limited number of participants. Accred Qual Assur 12:391–398. doi:10.1007/s00769-006-0247-0
58. Thompson M, Mathieson K, Owen L, Damant AP, Wood R (2008) The relationship between accreditation status and

performance in a proficiency test. Accred Qual Assur 14:73–78. doi:10.1007/s00769-008-0457-8

59. Heydorn K (2008) The application of data from proficiency testing to laboratory accreditation according to ISO 17025. Accred Qual Assur 13:703–709. doi:10.1007/s00769-008-0435-1
60. Vernelen K, Noble MA, Libeer J (2008) External quality assessment in microbiology: comparison of results from Belgian and Canadian laboratories with regard to their ability to identify Streptococcus pyogenes. Accred Qual Assur 13:501–504. doi:10.1007/s00769-008-0421-7
61. Rosario P, Martínez JL, Silván JM (2008) Comparison of different statistical methods for evaluation of proficiency test data. Accred Qual Assur 13:493–499. doi:10.1007/s00769-008-0413-7
62. Fisicaro P, Amarouche S, Lalere B, Labarraque G, Priel M (2008) Approaches to uncertainty evaluation based on proficiency testing schemes in chemical measurements. Accred Qual Assur 13:361–366. doi:10.1007/s00769-008-0402-x
63. Fisicaro P, Champion R, Labarraque G, Lalere B, Mace T, Sutour C (2008) Accred Qual Assur 13:367–372. doi:10.1007/s00769-008-0394-6
64. Chajduk E, Polkowska-Motrenko H, Dybczynski RS (2008) A definitive RNAA method for determination of selenium in biological samples: uncertainty evaluation and assessment of degree of accuracy. Accred Qual Assur 13:443–451. doi:10.1007/s00769-008-0377-7
65. Heydorn K (2008) Proficiency testing in the light of a new rationale in metrology. Accred Qual Assur 13:217–222. doi:10.1007/s00769-008-0368-8
66. Van Blerk M, Albarède S, Deom A, Gutiérrez G, Heller S, Nazor A et al (2008) Comparison of evaluation procedures used by European external quality assessment scheme organizers for haemoglobin concentration and leukocyte concentration. Accred Qual Assur 13:145–148. doi:10.1007/s00769-008-0364-z
67. Thompson M, Mathieson K, Damant AP, Wood R (2008) A general model for interlaboratory precision accounts for statistics from proficiency testing in food analysis. Accred Qual Assur 13:223–230. doi:10.1007/s00769-008-0356-z
68. Ellison SLR, Mathieson K (2008) Performance of uncertainty evaluation strategies in a food proficiency scheme. Accred Qual Assur 13:231–238. doi:10.1007/s00769-007-0353-7
69. Koch M, Baumeister F (2008) Traceable reference values for routine drinking water proficiency testing: first experiences. Accred Qual Assur 13:77–82. doi:10.1007/s00769-007-0349-3
70. Analytical Method Committee, The Royal Society of Chemistry (2010) The role of proficiency testing in method validation. Accred Qual Assur 15:73–79. doi:10.1007/s00769-009-0560-5
71. Ellison SLR (2009) Performance of MM-estimators on multimodal data shows potential for improvements in consensus value estimation. Accred Qual Assur 14:411–419. doi:10.1007/s00769-009-0571-2
72. Thomas A (2009) External Quality Assessment in laboratory medicine: is there a rationale to determine frequency of surveys? Accred Qual Assur 14:439–444. doi:10.1007/s00769-009-0563-2
73. Arnaud J, Jones RL, LeBlanc A, Lee M, Mazarrasa O, Parsons P et al (2009) Criteria to define the standard deviation for proficiency assessment for the determination of essential trace elements in serum: comparison of Z-scores based on the Horwitz function or on biological variability. Accred Qual Assur 14:427–430. doi:10.1007/s00769-009-0561-4
74. Belli M (2009) PT/EQA standards and guidelines: some remarks on quality and reliability of test items. Accred Qual Assur 14:421–426. doi:10.1007/s00769-009-0554-3
75. Priel M, Amarouche S, Fisicaro P (2009) Metrological traceability is not always a straight line. Accred Qual Assur 14:593–599. doi:10.1007/s00769-009-0540-9
76. Dubey V, Velikeloth S, Sliwakowski M, Mallard G (2009) Official proficiency tests of the organisation for the prohibition of chemical weapons: current status and future directions. Accred Qual Assur 14:431–437. doi:10.1007/s00769-009-0520-0
77. Priel M (2009) From GUM to alternative methods for measurement uncertainty evaluation. Accred Qual Assur 14:235–241. doi:10.1007/s00769-009-0518-7
78. Bashkansky E, Gadrich T, Kuselman I (2012) Interlaboratory comparison of test results of an ordinal or nominal binary property: analysis of variation. Accred Qual Assur 17:239–243. doi:10.1007/s00769-011-0856-0
79. Wong S (2011) Performance evaluation for proficiency testing with a limited number of participants. Accred Qual Assur 16:539–544. doi:10.1007/s00769-011-0816-8
80. Koch M, Magnusson B (2012) Use of characteristic functions derived from proficiency testing data to evaluate measurement uncertainties. Accred Qual Assur. doi:10.1007/s00769-012-0880-8
81. Koch M, Baumeister F (2012) On the use of consensus means as assigned values. Accred Qual Assur. doi:10.1007/s00769-012-0884-4
82. Ellison SLR, Hardcastle WA (2012) Causes of error in analytical chemistry: results of a web-based survey of proficiency testing participants. Accred Qual Assur. doi:10.1007/s00769-012-0894-2
83. Côté I, Robouch P, Robouch B, Bisson D, Gamache P, LeBlanc A et al (2012) Determination of the standard deviation for proficiency assessment from past participant's performances. Accred Qual Assur. doi:10.1007/s00769-012-0906-2
84. Jalukse L, Vabson V, Leito I (2006) In situ interlaboratory comparisons for dissolved oxygen concentration and Ph. Accred Qual Assur 10:562–564. doi:10.1007/s00769-005-0058-8
85. Thompson M (2006) Using mixture models for bump-hunting in the results of proficiency tests. Accred Qual Assur 10:501–505. doi:10.1007/s00769-005-0053-0
86. Stanton NV, Fritsch T, Geraghty C, Verostek MF, Weiner B, Parsons PJ (2006) The role of proficiency testing in the detection and resolution of calibration bias in the Lead Care R blood lead analyzer; limitations of peer-group assessment. Accred Qual Assur 11:590–592. doi:10.1007/s00769-006-0189-6
87. Gluschke M (2006) QUASIMEME results of laboratories involved in the German Marine Monitoring Programme for the North and Baltic Seas. Accred Qual Assur 11:470–473. doi:10.1007/s00769-006-0155-3
88. Van Berkel E, Boley N, Ellison S, Libeer J, Mäkinen I, Sandberg S (2006) Proficiency testing in analytical chemistry, microbiology and laboratory medicine—working group discussions on current status, problems and future directions. Accred Qual Assur 11:446–450. doi:10.1007/s00769-006-0111-2
89. Visser RG (2009) Reliability of proficiency test results for metals and phthalates in plastics. Accred Qual Assur 14:29–34. doi:10.1007/s00769-008-0465-8
90. Mäkinen I (2009) Use of uncertainty estimates as reported by participants in proficiency testing for the evaluation of their results: pros and cons. Accred Qual Assur 14:35–41. doi:10.1007/s00769-008-0462-y
91. Mitani Y, Lara-Manzano JV, Rodriguez-Lopez A (2008) Proficiency testing scheme for the harmonization and comparability of analytical measurements. Accred Qual Assur 13:421–426. doi:10.1007/s00769-008-0398-2
92. Gardner M (2008) Long-term proficiency testing for the UK Acid Waters Monitoring Network. Accred Qual Assur 13:255–260. doi:10.1007/s00769-008-0367-9
93. Tao L, Xiang H, Xu H, Sun J, Zhang M, Wang K et al (2010) Application of the FAPAS protocol to a proficiency testing on Pb and Cd extracted from glazed ceramic. Accred Qual Assur 15:155–162. doi:10.1007/s00769-009-0609-5

94. Drnovšek J, Pušnik I, Bojkovski J, Begeš G (2010) Regional cooperation in proficiency testing: strategy planning and practical experiences in Western Balkan countries. Accred Qual Assur 15:125–131. doi:10.1007/s00769-009-0569-9
95. Grimaldi M, Mancin M, Trevisan R, Mioni R (2009) Proficiency testing in food microbiology "Aqua": enumeration of Staphylococcus aureus. Accred Qual Assur 14:503–505. doi:10.1007/s00769-009-0572-1
96. De Souza GB, Nogueira ARA, Del-Santo VR, Picchi CMC, Guimarães ES, BarioniJr W (2009) Proficiency testing of animal nutrition laboratories. Accred Qual Assur 14:455–460. doi:10.1007/s00769-009-0562-3
97. Poulsen ME, Christensen HB, Herrmann SS (2009) Proficiency test on incurred and spiked pesticide residues in cereals. Accred Qual Assur 14:477–485. doi:10.1007/s00769-009-0555-2
98. Bayrak ÖT, Okandan E, Üçkardes H (2009) Inter-laboratory comparison scheme for fuel sector, LABKAR in Turkey. Accred Qual Assur 14:497–502. doi:10.1007/s00769-009-0542-7
99. Gaunt W, Whetton M (2009) Regular participation in proficiency testing provides long term improvements in laboratory performance: an assessment of data over time. Accred Qual Assur 14:449–454. doi:10.1007/s00769-009-0523-x
100. Cotman M, Drolc A, Tišler T (2009) Interlaboratory studies on wastewater toxicity using Daphnia magna. Accred Qual Assur 14:319–327. doi:10.1007/s00769-009-0519-6
101. Whetton M, Finch H (2009) Analytical performance is improved by regular participation in proficiency testing: an analysis of data from the Aquacheck proficiency testing scheme. Accred Qual Assur 14:445–448. doi:10.1007/s00769-009-0498-7
102. De Graaf DC, Ritter W, Jacobs FJ, Brunain M, Imberechts H, Mintiens K et al (2009) Lessons from the first international proficiency test for the detection of spores from the honey bee pathogen Paenibacillus larvae. Accred Qual Assur 14:273–276. doi:10.1007/s00769-009-0495-x
103. Drolc A, Cotman M (2009) Integration of metrological principles and performance evaluation in a proficiency testing scheme in support of the Council Directive 98/83/EC. Accred Qual Assur 14:199–205. doi:10.1007/s00769-009-0494-y
104. Urquiza MP, Mitani Y, Jaimes HON (2009) Proficiency tests to evaluate commercially available IVD kits for glucose and cholesterol measurements. Accred Qual Assur 14:269–271. doi:10.1007/s00769-009-0493-z
105. Mäkinen I, Hanste U (2009) Proficiency testing for measurement of radon (222Rn) in drinking water. Accred Qual Assur 14:473–476. doi:10.1007/s00769-009-0488-9
106. Antin L, Armishaw P (2010) Aspects of proficiency testing studies of trace elements in environmental samples with a focus on laboratory performance. Accred Qual Assur 15:467–471. doi:10.1007/s00769-010-0672-y
107. Kim Y, Lee J, Song K, Kang Y, Shin S, Jung G et al (2010) A review of proficiency testing scheme in Republic of Korea: in the field of water analysis. Accred Qual Assur 15:251–254. doi:10.1007/s00769-010-0635-3
108. Chung SWC, Tran JCH, Wong WW (2012) Capability of laboratories to determine core nutrients in foods: results of an international proficiency test. Accred Qual Assur 17:35–44. doi:10.1007/s00769-011-0833-7
109. Agbenin JO, Cantarella H (2011) Performance of commercial soil laboratories in a proficiency test program in Brazil. Accred Qual Assur 16:553–559. doi:10.1007/s00769-011-0814-x
110. Chelani AB, Moghe CA, Nimsadkar S, Gandhi K, Bodhe GL, Dhopte SM et al (2011) Evaluation of bias, precision, and systematic errors in proficiency testing of Cl2 and Cu concentration in water. Accred Qual Assur 16:379–382. doi:10.1007/s00769-011-0782-1
111. Pankratov I, Elhanany S, Henig S, Zaritsky S, Ostapenko I, Kuselman I (2010) Development of a proficiency testing scheme for a limited number of participants in the field of natural water analysis. Accred Qual Assur 15:459–466. doi:10.1007/s00769-010-0673-x
112. Aslan B, Gun-Munro J, Flynn GJ (2012) Role of proficiency testing in monitoring of standardization of hemoglobin A1c methods. Accred Qual Assur. doi:10.1007/s00769-012-0907-1
113. Whetton M, Baryla K, Finch H (2012) The effect of the choice of method for determining assigned value on the assessment of performance in melting point analysis. Accred Qual Assur. doi:10.1007/s00769-012-0908-0
114. Mancin M, Grimaldi M, Barco L, Trevisan R, Ruffa M, Mioni R et al (2012) Proficiency testing in food microbiology: experience from implementation of ISO/IEC 17043 and ISO/TS 22117. Accred Qual Assur. doi:10.1007/s00769-012-0904-4
115. Múrtula R, Soria E, Yáñez MA, Catalán V (2012) Proficiency testing schemes for the assessment of Legionella PCR methodologies. Accred Qual Assur. doi:10.1007/s00769-012-0903-5
116. Cordeiro F, Robouch P, Linsinger T, De La Calle B (2012) Determination of brominated flame retardants: a proficiency test. Accred Qual Assur. doi:10.1007/s00769-012-0900-8
117. Mbwambo K, Koch M (2012) Establishing PT schemes in developing countries: examples from Africa. Accred Qual Assur. doi:10.1007/s00769-012-0893-3
118. James VLA (2012) Evaluating participant performance in qualitative PT/EQA schemes. Accred Qual Assur. doi:10.1007/s00769-012-0895-1
119. Lehmann C (2012) Accrediting PT/EQA providers to ISO/IEC 17043. Accred Qual Assur. doi:10.1007/s00769-012-0892-4
120. Gun-Munro J (2012) The challenges and benefits of implementing the requirements of ISO/IEC 17043 by PT/EQA providers. Accred Qual Assur. doi:10.1007/s00769-012-0887-1
121. ISO 5725-1 (1994) Accuracy (trueness and precision) of measurement methods and results—part 1: general principles and definitions. International Organization for Standardization, Geneva, Switzerland
122. ISO 5725-2 (1994) Accuracy (trueness and precision) of measurement methods and results—part 2: basic method for the determination of repeatability and reproducibility of a standard measurement method. International Organization for Standardization, Geneva, Switzerland
123. ISO 5725-4 (1994) Accuracy (trueness and precision) of measurement methods and results—part 4: basic methods for the determination of the trueness of a standard measurement method. International Organization for Standardization, Geneva, Switzerland
124. ISO Guide 34 (2009) General requirements for the competence of reference material producers. International Organization for Standardization, Geneva, Switzerland
125. ISO 21748 (2010) Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation. International Organization for Standardization, Geneva, Switzerland
126. ISO/TS 22117(2010) Microbiology of food and animal feeding stuffs—specific requirements and guidance for proficiency testing by interlaboratory comparison. International Organization for Standardization, Geneva, Switzerland
127. ISO 13528 (2005) Statistical methods for use in proficiency testing by interlaboratory comparisons. International Organization for Standardization, Geneva, Switzerland
128. ISO/TS 20612(2007) Water quality—interlaboratory comparisons for proficiency testing of analytical chemistry laboratories. International Organization for Standardization, Geneva, Switzerland

129. ISO 15189 (2007) Medical laboratories—particular requirements for quality and competence. International Organization for Standardization, Geneva, Switzerland
130. ISO/IEC 17011 (2004) Conformity assessment—general requirements for accreditation bodies accrediting conformity assessment bodies. International Organization for Standardization, Geneva, Switzerland
131. ISO Guide 35(2006) Reference materials—general and statistical principles for certification. International Organization for Standardization, Geneva, Switzerland
132. ISO/TR 22971 (2005) Accuracy (trueness and precision) of measurement methods and results—practical guidance for the use of ISO 5725-2:1994 in designing, implementing and statistically analysing interlaboratory repeatability and reproducibility results. International Organization for Standardization, Geneva, Switzerland
133. ASTM E2489-11 Standard practice for statistical analysis of one-sample and two-sample interlaboratory proficiency testing programs. ASTM International, West Conshohocken, PA, USA
134. ASTM E2027-09 Standard practice for conducting proficiency tests in the chemical analysis of metals, ores, and related materials. ASTM International, West Conshohocken, PA, USA
135. ASTM D6674-01 Standard guide for proficiency test program for fabrics. ASTM International, West Conshohocken, PA, USA
136. ASTM D6792-07 Standard practice for quality system in petroleum products and lubricants testing laboratories. ASTM International, West Conshohocken, PA, USA
137. ASTM D7372-07 Standard guide for analysis and interpretation of proficiency test program results. ASTM International, West Conshohocken, PA, USA
138. IUPAC (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. Pure Appl Chem 78:145–196
139. IUPAC/CITAC Guide (2010) Selection and use of proficiency testing schemes for a limited number of participants—chemical analytical laboratories (IUPAC Technical Report). Pure Appl Chem 82:1099–1135
140. EUROLAB (2006) Technical Report 1/2006: guide to the evaluation of measurement uncertainty for quantitative results. http://www.eurolab.org/documents/1-2006.pdf. Accessed 3 Mar 2013
141. EUROLAB (2007) Technical Report: Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation. http://www.eurolab.org/documents/1-2007.pdf. Accessed 20 Mar 2013
142. NORDTEST (2003) Technical Report 537: handbook for calculation of measurement uncertainty in environmental laboratories. www.nordicinnovation.net/nordtest.cfm. Accessed 22 Mar 2013
143. BIPM, IEC, IFCC, ILAC, IUPAC, ISO, OIML (2008) Guide to the expression of uncertainty in measurement—GUM. JCGM 100, http://www.bipm.org/en/publications/guides/gum.html. Accessed 22 Mar 2013
144. EA-3/04 (2001) Use of proficiency testing as a tool for accreditation in testing. http://www.european-accreditation.org/publication/ea-3-04-g. Accessed 19 Mar 2013
145. EA-4/18 (2010) Guidance on the level and frequency of proficiency testing participation. http://www.european-accreditation.org/publication/ea-4-18-ta. Accessed 18 Jan 2013
146. ILAC (2010) P9: ilac policy for participation in proficiency testing activities. https://www.ilac.org/documents/ILAC_P9_11_2010.pdf. Accessed 11 Jan 2013
147. IAAC (2013) MD 038/13: application for the assessment and accreditation of proficiency testing providers. http://www.iaac.org.mx/Documents/Controlled/Mandatory/MD%20038%20ACCREDITATION%20OF%20PTP.pdf. Accessed 17 Mar 2014
148. APLAC (2008) PT 001. Calibration interlaboratory comparisons. Issue No. 5. http://www.aplac.org/documents/pt/aplac_pt_001_issue_5.pdf. Accessed 16 Jun 2014
149. APLAC (2010) PT 003. Proficiency Testing Directory. Issue No. 15. http://www.aplac.org/documents/pt/aplac_pt_003_issue_15.pdf. Accessed 16 Jun 2014
150. APLAC (2010) PT 005. Artefacts for Measurement Audits. Issue No. 2. http://www.aplac.org/documents/pt/aplac_pt_005_issue_2.pdf. Accessed 16 Jun 2014
151. APLAC (2010) PT 006. Proficiency Testing Frequency Benchmarks. Issue No. 2. http://www.aplac.org/documents/pt/aplac_pt_006_issue_2.pdf. Accessed 16 Jun 2014